IFW

PTO/SB/21 (09-04)
Approved for use through 07/31/2006. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

TRANSMITTAL FORM

(to be used for all correspondence after initial filing)

Total Number of Pages in This Submission

Application Number	10/781,352
Filing Date	February 17, 2004
First Named Inventor	Song ZHANG et al.
Art Unit	2641
Examiner Name	Unassigned
Attorney Docket Number	0331-049

ENCLOSURES *(Check all that apply)*

- ☐ Fee Transmittal Form
  - ☐ Fee Attached
- ☐ Amendment/Reply
  - ☐ After Final
  - ☐ Affidavits/declaration(s)
- ☐ Extension of Time Request
- ☐ Express Abandonment Request
- ☐ Information Disclosure Statement
- ☑ Certified Copy of Priority Document(s)
- ☐ Reply to Missing Parts/ Incomplete Application
  - ☐ Reply to Missing Parts under 37 CFR 1.52 or 1.53
- ☐ Drawing(s)
- ☐ Licensing-related Papers
- ☐ Petition
- ☐ Petition to Convert to a Provisional Application
- ☐ Power of Attorney, Revocation Change of Correspondence Address
- ☐ Terminal Disclaimer
- ☐ Request for Refund
- ☐ CD, Number of CD(s) _____
  - ☐ Landscape Table on CD
- ☐ After Allowance Communication to TC
- ☐ Appeal Communication to Board of Appeals and Interferences
- ☐ Appeal Communication to TC **(Appeal Notice, Brief, Reply Brief)**
- ☐ Proprietary Information
- ☐ Status Letter
- ☑ Other Enclosure(s) (please Identify below):

Assignment Recordation Form w/ authorization to charge $40 fee; executed Assignment; stamped return receipt postcard

Remarks

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT

Firm Name	Potomac Patent Group, PLLC		
Signature	S. M. duBois		
Printed name	Steven M. duBois		
Date	February 1, 2005	Reg. No.	35,023

CERTIFICATE OF TRANSMISSION/MAILING

I hereby certify that this correspondence is being facsimile transmitted to the USPTO or deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on the date shown below:

Signature	S. M. duBois		
Typed or printed name	Steven M. duBois	Date	February 1, 2005

Patent
Attorney's Docket No. 0331-049

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of	)	
	)	
Song ZHANG et al.	)	Group Art Unit: 2641
	)	
Application No.: 10/781,352	)	Examiner: Unassigned
	)	
Filed: February 17, 2004	)	
	)	
For: METHOD AND APPARATUS FOR DETECTING VOICE ACTIVITY	)	

SUBMISSION OF CERTIFIED PRIORITY DOCUMENT

Commissioner for Patents
Alexandria, VA 22313-1450

Sir:

Applicants claim priority of Canadian Patent Application No. 2,420,129, filed on February 17, 2003 and submit herewith a certified copy of the priority document.

Respectfully submitted,
POTOMAC PATENT GROUP PLLC

By: _____
Steven M. duBois
Registration No. 35,023

Date: February 1, 2005

Potomac Patent Group, PLLC
P.O. Box 270
Fredericksburg, VA 22404
(540) 361-1863

20510-49

Office de la propriété intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An Agency of Industry Canada

Bureau canadien des brevets
Certification

Canadian Patent Office
Certification

La présente atteste que les documents ci-joints, dont la liste figure ci-dessous, sont des copies authentiques des documents déposés au Bureau des brevets.

This is to certify that the documents attached hereto and identified below are true copies of the documents on file in the Patent Office.

Specification and Drawings, as originally filed, with Application for Patent Serial No: **2,420,129**, on February 17, 2003, by **CATENA NETWORKS, CANADA INC.,** assignee of Song Zhang and Eric Verreault, for "A Method for Robustly Detecting Voice Activity.

CERTIFIED COPY OF
PRIORITY DOCUMENT

Agent certificateur/Certifying Officer

March 10, 2004
Date

Canada

(CIPO 68)
04-09-02

OPIC CIPO

A METHOD FOR ROBUSTLY DETECTING VOICE ACTIVITY

Background of Invention:

Voice activity detection (VAD) techniques have been widely used in digital voice communications to reduce voice data rate to achieve either spectral efficient voice transmission or power efficient voice transmission for wireless devices. The essential part of VAD algorithms is to effectively distinguish voice signal and background noise signal, where multiple aspects of signal characteristics, like energy level, spectral contents, periodicity and stationarity, etc., have to be explored. Traditional VAD algorithms tend to use heuristic approaches to apply some limited subset of the characteristics to detect voice presence, which, in practice, are very difficult to achieve high voice detection rate and low false alarm rate due to the heuristic nature of the technique. To address the performance issue of heuristic algorithms, more sophisticated algorithms are developed to simultaneously monitor multiple signal characteristics and try to make a detection decision based on some joint metrics. These algorithms do demonstrate good performance, but at the same time, they often lead to complicated implementations or inevitably become an integrated component of some specific voice encoder algorithm. Lately, a statistical model based VAD algorithm is studied and shows good performance and simple mathematical framework [1]. The challenge, however, to make this new algorithm practical to effectively estimate both voice and noise signal power on each frequency component.

Detailed Description of invention

The invention disclosed here describes a robust statistical model based VAD algorithm, which does not rely on any presumptions of voice and noise statistical characters and can quickly train itself to effectively detect voice signal with good performance. What makes it more attractive is that it works as a stand-alone module and is independent of the type of voice encoders.

The key advantages of this method are:

a. Use statistical model based approach with proven performance and simplicity.
b. Self-training and adapting without reliance on any presumptions of voice and noise statistical characters.
c. An adaptive detection threshold that makes the algorithm work in any signal-to-noise ratio (SNR) scenarios.
d. A generic stand-alone structure that can work with different voice encoders.

1.Mathematical Framework

The underlying mathematical framework for the algorithm is the log likelihood ratio of the event when there is noise only and the event when there are both voice and noise. It can be mathematically formulated as:

Let $y(t) = x(t) + n(t)$ be a frame of received signal and **Y** be its corresponding pre-selected set of complex frequency components. Further, two events are defined as:

Y = N, as H_0 -- speech absent,
Y = X + N, as H_1 – speech present,

Where, **X** and **N** are corresponding pre-selected set of complex frequency components of voice `x(t)` and `n(t)` respectively. It is sufficiently accurate to model **Y** as a jointly Gaussian distributed random vector with each individual component as an independent complex Gaussian variable, and **Y**'s PDF conditioned on H_0 and H_1 can be expressed as:

$$p(Y \mid H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi \lambda_N(k)} \exp\left(-\frac{|Y_k|^2}{\lambda_N(k)} \right)$$

$$p(Y \mid H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi [\lambda_X(k) + \lambda_N(k)]} \exp\left(-\frac{|Y_k|^2}{[\lambda_X(k) + \lambda_N(k)]} \right)$$

where, $\lambda_X(k)$ and $\lambda_N(k)$ are the variances of the voice complex frequency component X_k and the noise complex frequency component N_k respectively.

Let log likelihood ratio (LLR) of the kth frequency component be defined as:

$$\log(\Lambda_k) = \log(\frac{p(Y_k \mid H_1)}{p(Y_k \mid H_0)}) = \left(\frac{\gamma_k \cdot \xi_k}{1 + \xi_k} \right) - \log(1 + \xi_k)$$

where, ξ_k and γ_k are the so-called a priori signal-to-noise ratio (pri-SNR) and a posteriori signal-to-noise ratios (post-SNR) respectively, as defined:

$$\xi_k = \frac{\lambda_X(k)}{\lambda_N(k)}$$

$$\gamma_k = \frac{|Y_k|^2}{\lambda_N(k)}$$

Then, the LLR of vector **Y** given H_0 and H_1, which is what a VAD decision based on, can expressed as:

$$\log(\Lambda) = \sum_k \log(\Lambda_k) = \sum_k \log(\frac{p(Y_k \mid H_1)}{p(Y_k \mid H_0)}) = \sum_k (\left(\frac{\gamma_k \cdot \xi_k}{1 + \xi_k} \right) - \log(1 + \xi_k))$$

A LLR threshold developed based on SNR level can be used to make a decision on if voice signal is present or not.

2. Basic Operations

The general flow of the algorithm is illustrated in Figure 1, and each function block is explained in details as follows:

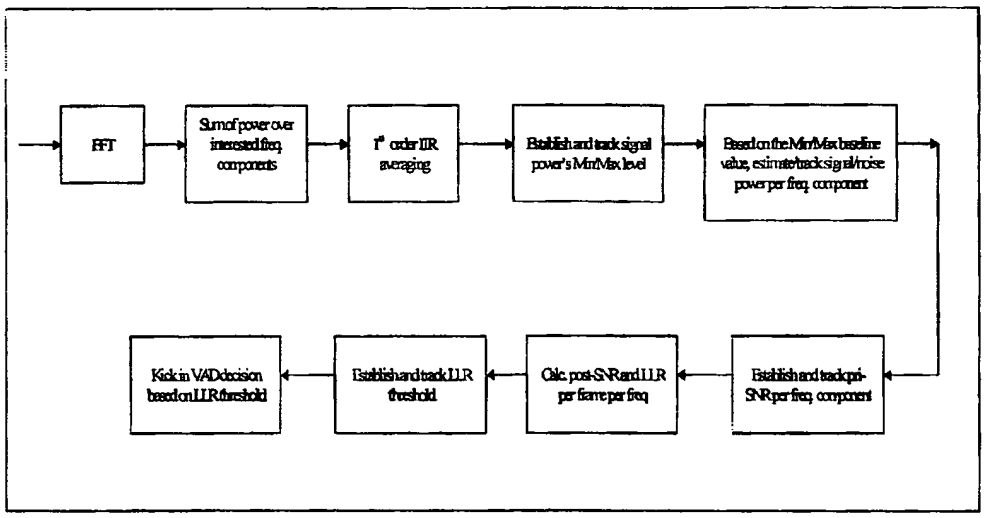

Figure 1 Flow diagram of VAD algorithm

1. For a inbound 5-ms signal frame of 40 samples, 32/64-point FFT is performed. If 32-point FFT is performed, 40-sample frame is truncated to 32 samples. In the case of 64-point FFT, 40-sample frame is zero padded.

Note: inbound signal frame size and FFT size can change depending on the implementation.

2. From FFT output, sum of signal power over pre-selected frequency set is calculated and go through a 1st-order IIR averager to extract long-term signal dynamics, as illustrated in Figure 2 and Figure 3. IIR averager's forgetting factor is chosen such that signal's peaks and valleys are kept.

Reference:

[1] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A Statistical Model-Based Voice Activity Detection," IEEE Signal Processing Letters, Vol. 6, No. 1, Jan. 1999.

Claims

This invention disclosure claims the following:

1) The method to use the statistical model based mathematical formulation to do VAD.
2) The method to estimate and track voice signal and noise signal power in the frequency domain.
3) The method to establish and adapt the LLR threshold for VAD detection.

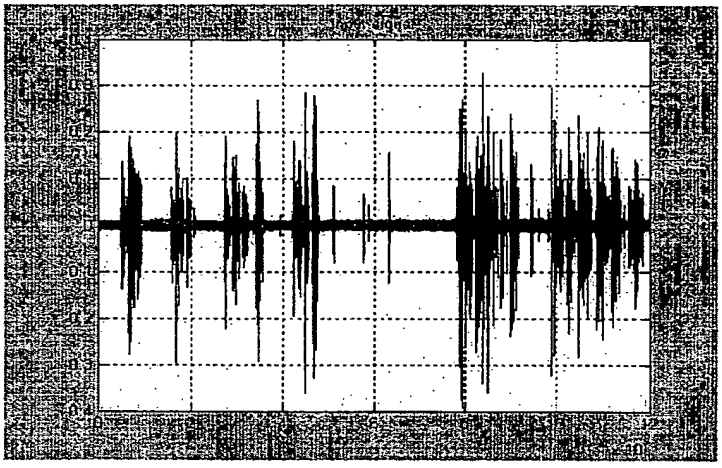

Figure 2 Noise corrupted voice signal

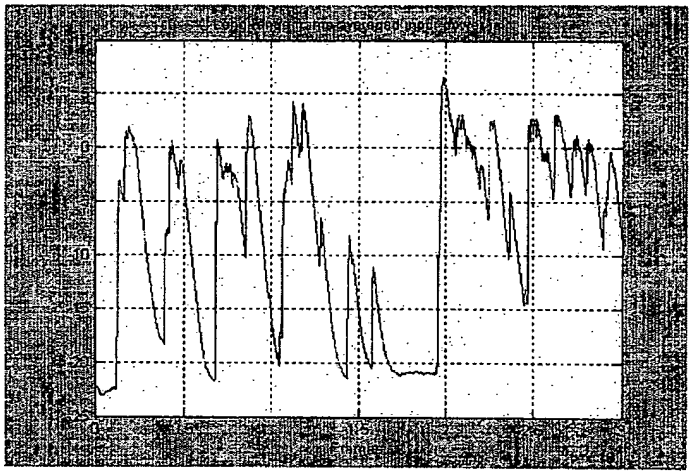

Figure 3 Signal dynamics after IIR averaging

3. As signal power's dynamic is available, based on a pre-configured min/max signal level gap threshold, say 12dB, initial min/max level can be established using simple absolute level detector. Afterwards, a slow 1st-order averager is used to slowly update two levels to follow signal's dynamic change based on pre-defined margin value. To build in high level of system stability to prevent min/max gap collapse, min level adaptation is designed such that it is quicker to adapt down than adapt up. Similar treatment is done on the max level adaptation as well. In the case the gap does collapse, the system is reset to re-establish valid min/max baseline. Figure 4 Illustrates what the min/max level looks like.

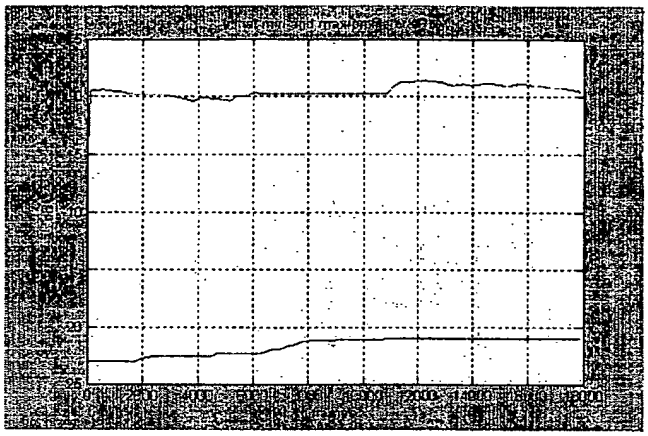

Figure 4 Establishment and tracking of min/max level

4. Using the slowly-moving min/max levels as a baseline, the algorithm defines a range of signal to be considered as noise and voice respectively, and a 1st-order IIR averager is used to calculate noise power and voice power respectively. The establishment of noise and voice power is illustrated in Figure 5 and Figure 6.

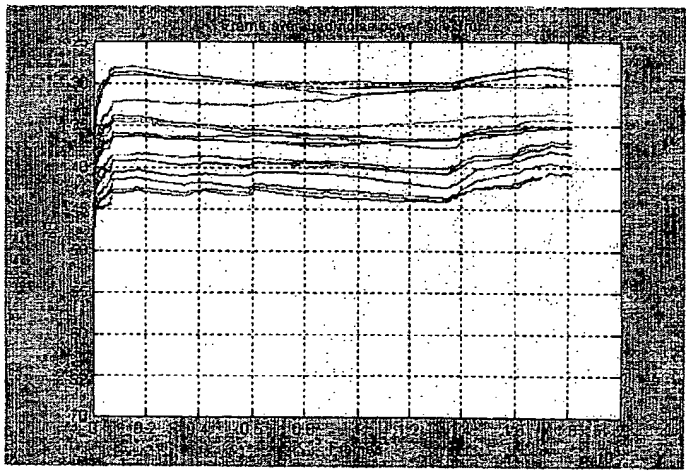

Figure 5 Establishment of noise power profile

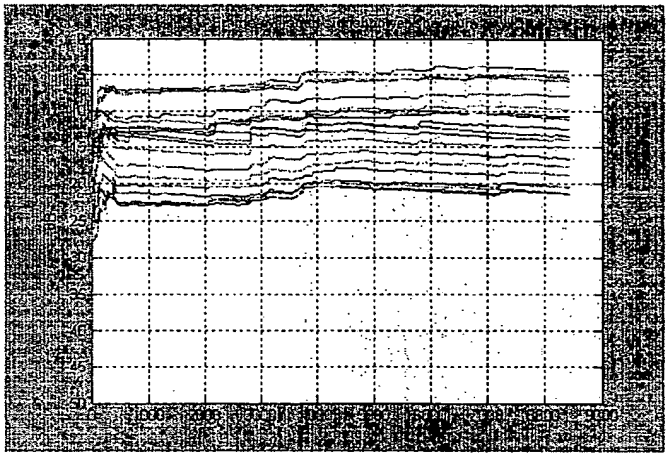

Figure 6 Establishment of voice power profile

5. After both noise power and voice power are established, a pri-SNR profile against the frequency component set can be calculated and tracked, again, using a 1st-order IIR averager. The result is shown in Figure 7.

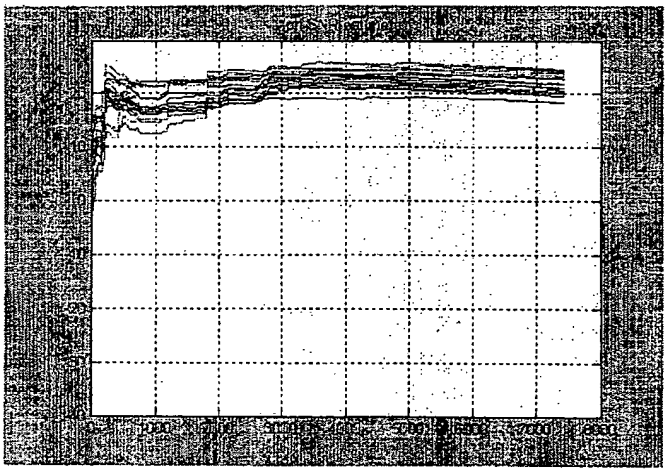

Figure 7 Establishment and tracking of pri-SNR profile

6. After the pri-SNR profile is available, the corresponding post-SNR profile and LLR profile can be calculated on a frame-by-frame basis. With the availability of LLRs over time and the knowledge of what is considered as noise frames from step 4, LLR threshold can be established and tracked using a 1st-order IIR averager. LLR distribution along the time and adaptation of LLR threshold are illustrated in Figure 8 and Figure 9.

Note: What is considered as noise frames in step 4 is not reliable enough for VAD purpose, as shown in Figure 9, where some of the LLR values are well above zero.

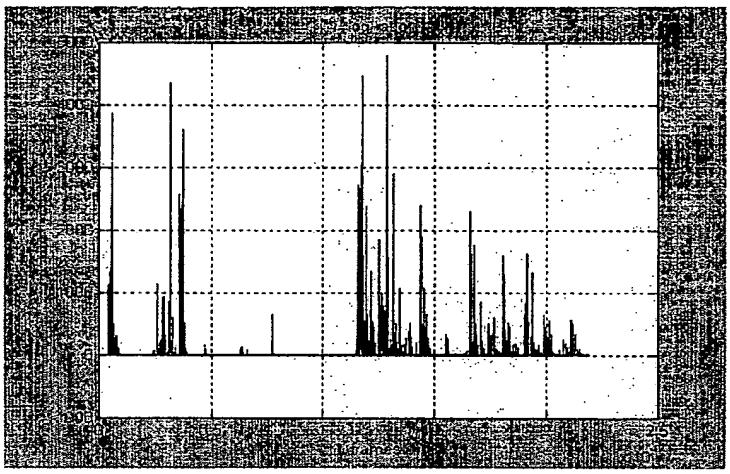

Figure 8 LLR distribution over time

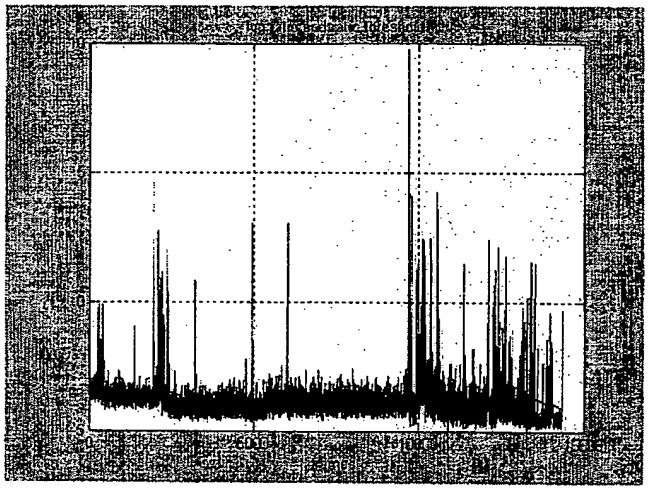

Figure 9 LLRs of as considered noise frames and LLR threshold adaptation

7. After the LLR threshold is available, silence detection is kicked in on a frame-by-frame basis. A frame is considered as silence if its LLR is below LLR threshold + x dB of margin and silence suppression is not triggered unless there are x number of consecutive silence frames (hang-over time). Figure 10 shows noise-removed voice signal.

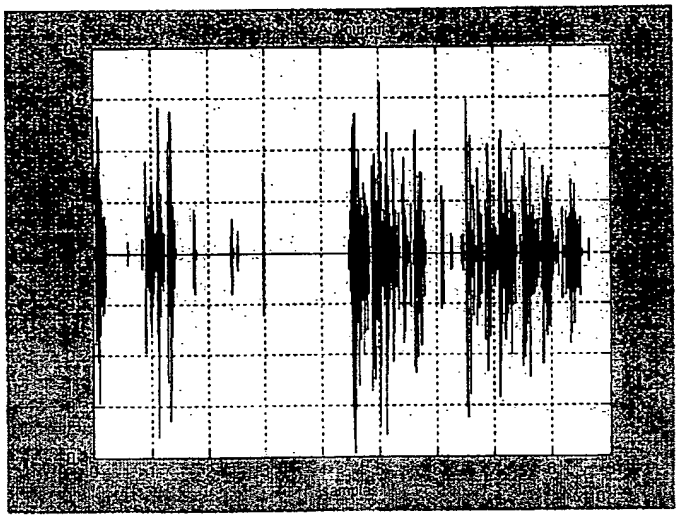

Figure 10 Noise suppressed voice signal